Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality

David L. Donoho
Department of Statistics
Stanford University

August 8, 2000

Abstract

The coming century is surely the century of *data*. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. Hyperspectral Imagery, Internet Portals, Financial tick-by-tick data, and DNA Microarrays are just a few of the better-known sources, feeding data in torrential streams into scientific and business databases worldwide.

In traditional statistical data analysis, we think of *observations* of instances of particular phenomena (e.g. instance ↔ human being), these observations being a vector of values we measured on several *variables* (e.g. blood pressure, weight, height, ...). In traditional statistical methodology, we assumed many observations and a few, well-chosen variables. The trend today is towards more observations but even more so, to radically larger numbers of variables – voracious, automatic, systematic collection of hyper-informative detail about each observed instance. We are seeing examples where the observations gathered on individual instances are curves, or spectra, or images, or even movies, so that a single observation has dimensions in the thousands or billions, while there are only tens or hundreds of instances available for study. Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector. We can say with complete confidence that in the coming century, high-dimensional data analysis will be a very significant activity, and completely new methods of high-dimensional data analysis will be developed; we just don't know what they are yet.

Mathematicians are ideally prepared for appreciating the abstract issues involved in finding patterns in such high-dimensional data. Two of the most influential principles in the coming century will be principles originally discovered and cultivated by mathematicians: the blessings of dimensionality and the curse of dimensionality.

The curse of dimensionality is a phrase used by several subfields in the mathematical sciences; I use it here to refer to the apparent intractability of systematically searching through a high-dimensional space, the apparent intractability of accurately approximating a general high-dimensional function, the apparent intractability of integrating a high-dimensional function.

The blessings of dimensionality are less widely noted, but they include the concentration of measure phenomenon (so-called in the geometry of Banach spaces), which means that certain random fluctuations are very well controlled in high dimensions and the success of asymptotic methods, used widely in mathematical statistics and statistical physics, which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated.

There is a large body of interesting work going on in the mathematical sciences, both to attack the curse of dimensionality in specific ways, and to extend the benefits of dimensionality. I will mention work in high-dimensional approximation theory, in probability theory, and in mathematical statistics. I expect to see in the coming decades many further mathematical elaborations to our inventory of Blessings and Curses, and I expect such contributions to have a broad impact on society's ability to extract meaning from the massive datasets it has decided to compile.

At the end of my talk, I will also draw on my personal research experiences. This suggest to me (1) ongoing developments in high-dimensional data analysis may lead mathematicians to study new problems in for example harmonic analysis; and (2) that many of the problems of low dimensional data analysis are unsolved and are similar to problems in harmonic analysis which have only recently been attacked, and for which only the merest beginnings have been made. Both fields can progress together.

Dedication. To the Memory of John Wilder Tukey 1915-2000.

Keywords. Data Mining. Multivariate Data Analysis. Principal Components Analysis. Independent Components Analysis. Computational Harmonic Analysis. Randomized Algorithms.

Acknowledgments. The effort to prepare this expository work was supported by NSF grant DMS 95-05152, and by NSF grant DMS 98-72890.

The following colleagues/collaborators have helped with time, discussions, ideas, figures, citations.

Amir Averbuch	Tel-Aviv University
Raphy Coifman	Yale University
Jerome Friedman	Stanford University
Trevor Hastie	Stanford University
Iain Johnstone	Stanford University
Art Owen	Stanford University
David Scott	Rice University
Bob Stine	University of Pennsylvania
Robert Tibshirani	Stanford University
Walter Willinger	ATT Shannon Labs

The following scientists have helped with time, leads, and presentation materials.

Gregory Piatetsky-Shapiro	XChange
Edward Bosch	U.S. Army Topographic Eng. Ctr.
John Elder	Datamininglab.com
Al Inselberg	Tel Aviv University
Tzyy-Ping Jung	Salk Institute
Terry Sejnowski	Salk Institute
Jean-Luc Starck	Centre Europeen d'Astronomie
Arne Stoschek	Stanford University
Peter Weinberger	Renaissance Technologies

I am greatly indebted to all of the above individuals. I would also like to thank Dr. Maureen Clerc, Dr. Ana Georgina Flesia, Kris Jennings, Ofer Levi, and Dev Millstein for editorial help in preparing this talk.

1 Introduction

1.1 August 8, 2000

The morning of August 8, 1900, David Hilbert gave an address at the International Congress of mathematicians in Paris, entitled ‘Mathematical Problems’ [50, 24]. Despite the fact that this was not a plenary address and that it was delivered as part of the relatively non-prestigious section of History and Philosophy of Mathematics, Hilbert’s lecture eventually came to be considered as the most important event of that Congress. In that lecture, Hilbert laid out 23 problems or problem areas which he had identified as important for the future development of mathematics; over the ensuing decades, the problem list attracted a great deal of attention and many of the problems were attacked energetically and even solved. In fact the lecture is very readable – see [27] for English translations of the article that Hilbert later published.

Today, the morning of August 8, 2000, we are gathered together exactly a century after Hilbert’s address, to consider the Mathematical Challenges of the Twenty-First century. Other talks at this conference cover a wide range of fascinating and deep problems, and are to be delivered by eminent and supremely talented mathematicians of towering achievements. This conference is in a very obvious sense an homage to Hilbert’s inspired idea. And it is not the only one; I am aware of other conferences organized under the same spell.

1.2 Hot Air?

Presumably we are here today because we take the topic seriously, and we think that what we say might in fact, like the Hilbert problems, have some bearing on, or at the very least, weak correlation with, the development of mathematics in the coming century.

There is a risk of heaviness and pretentiousness in such an undertaking. To reduce this, I have chosen a perhaps peculiar topic and narrative line for this presentation.

I’d like to focus on a set of issues I have learned about over the last quarter century as I progressed through my scientific career. I am absolutely convinced that these themes I will raise represent major societal and scientific trends in the coming century, that mathematics has had significant contributions to these areas in the past, that it can play a significant role in the future, that it would be valuable to the audience to be reminded of these trends, and that most of the audience will find something to ‘latch onto’ in what I say. I do not report here existing results of awesome mathematical depth, certainly nothing that could compete with the intricate issues being discussed here in connection with famous and longstanding problems of topology, analysis and algebra. Perhaps one of my messages is that should mathematics provide such results, there would be major impacts, and that I think results in that direction are possible. However, another of my messages is that the apparent superficiality of some of these ideas is itself worth taking note of, and presenting another kind of challenge for mathematics in the coming century.

1.3 John Tukey

As I was preparing this lecture, I received the sad news that John Wilder Tukey, Professor Emeritus at Princeton, had had a stroke. A few weeks later I received the even more distressing news that he had passed away. His obituary appeared July 28, 2000 in the *New York Times.*

John Tukey was my undergraduate Thesis adviser at Princeton, and had a great deal to do with my intellectual development in the crucial years from age 17 to 21 - both through his personal interactions with me, and through his writings and impact on the Statistics profession This is my first lecture since his passing, and I am dedicating it to John's memory. This is particularly fitting since, as we shall see, John's interests and life story are intertwined with the themes I am going to present.

Many in the audience will know of Tukey's more visible distinctions. He coined the words 'Software' and 'Bit', creating a lasting contribution to the English language; he and collaborators discovered two FFT algorithms and thereby fomented a revolution in signal processing and applied mathematics. Some in the audience will know more than these basics; in fact Tukey made fundamental contributions in many areas of the mathematical sciences. Working in topology in the 1930's, he invented filters and ultrafilters before Bourbaki, but chose the unfortunate name 'Phalanx'; he contributed a form of the Axiom of Choice, now named after him; he discovered basic phenomena in connection with stable laws in probability theory, with nonparametric spectral density estimation, and with higher-order statistics. He initiated major research fields in statistics – multiple comparisons and robust statistical methods. He was also a prophet, recognizing important issues decades before their time, for example the importance of non-Gaussianity in data and non-Linearity in signal processing.

In this talk I will not make an attempt to do real justice to the breadth and extent of his contributions. His abilities were recognized, both with impressive posts at Princeton and at Bell Labs, with membership of key science policy committees at the National Research Committee, with the IEEE Medal – a substantial distinction that few other mathematical scientists (if any) have been granted – and with the National Medal of Science.

It is important for the audience to know that John was very idiosyncratic, having been home schooled, and so had unusual means of self-expression; he also drew on an unusually broad array of knowledge (he was trained as a chemist before entering mathematics). He invented words left and right, many of which were not as apposite as 'bit'. John often purposely mangled words, for example dyslexing spectrum to make 'cepstrum', and he also invoked action phrases in place of established (numbing) terminology, using the overly evocative term 'head-banger' to describe a thresholding operation, and so on – the examples could go on endlessly. Talking to John often required a lengthy initiation into a new terminology that he was trying out at the time, based on recent thinking of his – some times about things he'd only been thinking about that day. I once worked as John's interpreter, accompanying Tukey to an industrial research lab, (I hope) explaining what he had in mind to a group of conventionally trained scientists.

The essayist Lionel Trilling has defined a Genius as one who exhibits a very strong personality, to the point of being a 'type', which gets expressed thoroughly in their entire intellectual output [60]. An important point for Trilling is that the Genius is consistently driven to express his uniqueness of vision and personality, whether or not they are correct or appropriate to the situation at hand. Trilling contrasted this with a kind of anti-genius, an intellectual with no distinctive vision or personality, but who struggles with the situation at hand rather than with the expression of deeper or broader drives and intuitions. John was indeed a Genius in Trilling's sense; he sensed many things far ahead of his time, and his unique cast of mind and expression helped him to bring up issues that could not otherwise have found expression in the mathematical sciences of his time. Those close to him recognize the greatness of this aspect of his personality and biography, a greatness which is hard to provide traditional recognition.

On the other hand, all should recognize that despite his status as Bona Fide Genius, not all John's ideas led to happy successes, even those in which he made great investments.

1.4 Data Analysis

A major initiative of Tukey's, at the peak of his eminence in the 1960's, was to recognize *Data Analysis* as an emerging activity/discipline, one distinct from *Mathematical Statistics*, and requiring its own literature.

This was in some ways a strange initiative. In the 1960's Tukey founded robust statistics (robust linear regression, contaminated normal distribution, 1962) nonlinear signal processing (cepstrum 1962), fast Fourier transforms (1964). Simply elaborating his many ideas in these directions could have led to an honored and very rewarding career. However, in 1962, in a major Plenary address at the annual meeting of the Institute of Mathematical Statistics, entitled 'The Future of Data Analysis', he began a provocative initiative which he maintained for the next 15 years, through the publication of 'Exploratory Data Analysis'.

The initiative was important because in some ways it functioned for statistics like the Hilbert Problems, being widely cited, and inspiring many young persons, myself included. The initiative was however, not greeted with universal approval; in fact it was viewed as scandalous in some quarters [65]. The 'Future of Data Analysis' has, in the end, functioned as a kind of Hilbert Problems in reverse, arguing that a certain kind of de-emphasis of mathematical problem solving in statistics is called for, in fact an abandonment of many of the traditional precepts by which intellectual life in academic statistics had been developing. Tukey's obituary in the *Times* chose to emphasize, not his serious mathematical achievements, but exactly this aspect: a respected mathematician turning his back on proof, focusing on analyzing data rather than proving theorems. For example, statistician Howard Wainer is quoted by the *Times* on this recusal:

> "He legitimized that, because he wasn't doing it because he wasn't good at math," Mr. Wainer said. "He was doing it because it was the right thing to do."

So what was Tukey's vision in the early 1960's? Roughly speaking, and I do not cite chapter and verse here, not knowing if I can produce a specific citation, we can say that Tukey made the points: Data are here now, they'll be coming on more and more in the future, we must analyze them, often using very humble means, and insistence on mathematics – for example on deep mathematical analysis and proof – will likely distract us from recognizing these fundamental points.

In the article "Data Analysis, including Statistics" (1968) written with Fred Mosteller, Tukey made the polemical point that data analysis was a potentially huge field, into which statistics – with its grounding as a subdiscipline of the mathematical sciences, via probability theory, decision theory, and analysis – fit only as a small segment. In that article, Mosteller and Tukey called attention to new sources of data – such as Satellites – automatically generating huge volumes of data, whose assimilation called for a wide variety of data processing and analysis tools. In that telling, the traditional ideas of mathematical statistics, of inferential ideas such as hypothesis testing and confidence statements, had relatively little to offer, but there was an enormous amount to be done, and much of it could be done with rather homely means.

In other articles, Tukey also tried to emphasize the separateness of data analysis from mathematics. I remember well some of the basic themes: data analysis is an activity all its own, a kind of lifelong avocation; one does not need to be a mathematician to be a data analyst – in fact there is no connection, one could as well be a biologist, etc.

There was initial resistance to Tukey's message in the community of mathematical statisticians; in fact some objected that Tukey's original address on the 'Future of Data Analysis' had no business being presented in such a forum of mathematical scientists [65]. On the other hand, a data analysis community crystallized around Tukey's conception, and its descendant communities are visible today, a substantial body of academic and industrial statisticians that emphasize data analysis over mathematical analysis and proof.

1.5 Data Analysis Today

Tukey was certainly correct about many things in taking up this initiative. Today, the world is awash in data, as I will describe below, and there are many more data analysts than there are statisticians; conferences in genomics, in remote sensing, in financial engineering, and other fields together tens of thousands of professionals to analyze data. The size of the mathematical statistics profession is very small next to the size of the data analysis profession. And data analysis is growing at a heady rate.

The ensuing patterns of thought are quite distinct from mathematics. The practitioners use some mathematical concepts in their work, but mathematical analysis is not heavily used, and theorems are not considered important. Part of this is because, owing to computer simulation, we no longer need theorems as much in day-to-day data analysis, but also because the cultural orientation in these fields places little emphasis on the theorem as a valuable cultural artifact. This last feature can also can be identified with Tukey, as mentioned in the quote from his *Times* Obituary cited earlier.

1.6 Break From the Past

Let me emphasize just how distinct from the mathematical tradition the data analysis community has become. Each year there are dozens of conferences on topics ranging from genomics to satellite imagery to chemometrics, in which the scientific *modus operandi* is: propose a new kind of phenomenon or a new kind of artifact in data, suggest a processing strategy that on heuristic grounds should ameliorate it, and report a computational example or two which shows that the strategy has at least a few apparent successes. There is often

- No formal definition of the phenomenon or artifact, in terms of a carefully stated mathematical model.
- No formal derivation of the proposed processing strategy, suggesting that it is in some sense naturally associated with the phenomenon to be treated.
- No formal analysis justifying the apparent improvement in simulations

This is a far cry from the intellectual traditional of mathematics.

1.7 Statistics and Mathematics

And of statistics. For some in the audience, it will be important to recall that things have not always been as they are now. Before the development of the 'data analysis' movement, statistics was closely tied to mathematics and theorems. Many of the most commonplace tools in everyday data analysis were created by mathematically talented individuals and supported by mathematically inspired standards of argument.

In fact, Gauss and Laplace made lasting contributions to statistical data analysis. Fourier was supported, at his low point following Napoleon's banishment, by a sinecure at an institute of Statistics. In the 1930's, Kolmogorov and Wiener both made substantial contributions to the foundation of statistical methodology.

There is also a tradition going in the other direction, of statisticians stimulating mathematical developments. In the twentieth century, we can mention Harold Hotelling, who originated the method of Principal Components, and who stimulated the well-known work of Hermann Weyl on the Volume of Tubes. Hotelling gave a talk on a statistical problem at Princeton which attracted Weyl's interest; out of this grew simultaneous publication in a single issue of *Amer. J. Math.* of a statistical paper by Hotelling [29] and a differential-geometric paper by Weyl [68]. Out of this grew a substantial amount of modern differential geometry.

Even if we focus attention on the basic tools of modern data analysis, from regression to principal components, we find they were developed by scientists working squarely in the mathematical tradition, and are based on theorems and analysis. Finally, let me say that there is a quite active and vigorous community of mathematical statisticians!

1.8 Far Enough Down That Road

I would now like to propose that Tukey's insight has run its course. Undoubtedly Tukey was right to emphasize the need for a schism, in which Data Analysis would form a separate culture from Mathematical Statistics. Over the last forty years, Data Analysis has developed at breakneck pace, responding to the rapid advances in information technology: massive data storage, rapid throughput, effective algorithms for basic problems. Over the last twenty years particularly, the largest contributions of statistics to data analysis have been in the formalization of information technology for data analysis, in the form of software packages like S and S-Plus, CART, Data Desk, GLIM, MacSpin, and in the identification of ways we can use information technology to substitute for analysis – bootstrap, Monte-Carlo Markov Chain.

However, this trend is, I think, now complete; it is highly unlikely that further developments in information technology will do much to solve any of the existing important structural problems for data analysis. Moreover, those fundamental problems are omnipresent, and now that the Data Analysis movement has insinuated itself into every branch of society, they affect many fields simultaneously. And the missing ingredient in facing those problems, it seems to me, is mathematics.

1.9 This Talk

In this talk I will reconsider data analysis and its relation to the mathematical sciences. I will suggest that, while over the last forty years, data analysis has developed by separating itself from its mathematical parent, there is now need for a reconnection, which might yield benefits both for mathematics and for data analysis.

We are now in a setting where many very important data analysis problems are high-dimensional. Many of these high-dimensional data analysis problems require new or different mathematics. A central issue is the curse of dimensionality, which has ubiquitous effects throughout the sciences. This is countervailed by three blessings of dimensionality. Coping with the curse and exploiting the blessings are centrally mathematical issues, and only can be attacked by mathematical means.

I will finish on a personal note, pointing out some recent research experiences which suggest to me that modern data analysis has actually reached a point that it can pose stimulating new questions to established mathematical disciplines such as harmonic analysis.

2 Data

Over the last few decades, data, data management, and data processing have become ubiquitous factors in modern life and work. In this section, I will remind the audience of a few indicators of a major societal trend.

2.1 Recent Data Trends

Huge investments have been made in various data gathering and data processing mechanisms. The information technology industry is the fastest growing and most lucrative segment of the world economy, and much of the growth occurs in the development, management, and warehousing of prodigious streams of data for scientific, medical, engineering, and commercial purposes. Some recent examples include:

- Biotech Data. Virtually everyone is aware of the fantastic progress made in the last five years in gathering data about the human genome. A common sight in the press is pictures of vast warehouses filled with genome sequencing machines working night and day, or vast warehouses of compute servers working night and day, as part of this heroic effort [53].

  This is actually just the opening round in a long series of developments. The genome is only indirectly related to protein function and protein function are only indirectly related to overall cell function. Over time, the focus of attention will go from genomics to proteomics and beyond. At each round, more and more massive databases will be compiled.

- Financial Data. Over the last decade, high-frequency financial data have become available; in the early to mid 1990's data on individual currency trades, became available, tracking individual transactions. Now with the advent of new exchanges such as Island.com, one can obtain individual bids to buy and sell, and the full distribution of such bids.

- Satellite Imagery. Providers of satellite imagery have available a vast database of such images, and hence N in the millions. Projects are in place to compile databases to resolve the entire surface of the earth to 1 meter accuracy. Applications of such imagery include natural resource discovery and agriculture.

- Hyperspectral Imagery. It is now becoming common, both in airborne photographic imagery and satellite imagery to use hyperspectral cameras which record, instead of three color bands RGB, thousands of different spectral bands. Such imagery is presumably able to reveal subtle information about chemical composition and is potentially very helpful in determining crop identity, spread of diseases in crops, in understanding the effects of droughts and pests, and so on. In the future we can expect hyperspectral cameras to be useful in food inspection, medical examination, and so on.

- Consumer Financial Data. Every transaction we make on the web, whether a visit, a search, a purchase, is being recorded, correlated, compiled into databases, and sold and resold, as advertisers scramble to correlate consumer actions with pockets of demand for various goods and services.

2.2 Extrapolation of Trends

The existing trends are likely to accelerate in the future, as each year new sensors, sources of data are invented, we scale to higher densities. A very important additional trend is that society will more and more think of itself as a data-driven society, a consumer of data.

It is becoming accepted that there is a data industry, that firms devoted to the creation and management of data – for example biotech companies – can be as valuable as firms creating physical tangible objects.

It is also becoming accepted that consumers will agree to become data processors. For example, a few years ago, a 1024*1024 image was considered quite a substantial object for handling on a modern computer, and only computer scientists were really working with digital imagery. Now, consumer cameras costing a few hundreds of dollars, generate such images routinely. Consumers are becoming familiar with the process of capturing images, downloading them onto their home computers, processing them with various software tools, creating custom imagery. Such consumer acceptance will doubtless fuel further investment and technological development.

2.3 Ubiquity of Data Supply

Another important trend is the open availability of data, for example over the internet. We see this everywhere; I will mention just two striking examples:

- `www.Island.com` makes available to everyone information of unprecedented detail about the functioning of a public stock exchange. A user with a web browser can obtain attractive presentations of all pending bids to buy and sell a certain stock at a certain time.

- `www.DoubleTwist.com` makes available annotated genomic data over the web to paying corporate customers. One gets not only the genome, but also a wealth of information correlated to it.

2.4 Ubiquity of Data Demand

We can envision a future in which data becomes as widely consumed as entertainment.

In his book *Mirror Worlds*, David Gelertner proposes a vision of a world-to-arrive-real-soon-now in which data and computing are ubiquitous, and a new kind of relationship of people to data emerges. In this new world, measurements and compilations about essentially any conceivable quantities are freely available over the network, to people in their homes (for example); data could include everything conceivable. For those interested in their locality, they could tap into data ranging from municipal expenditures by category to current automobile traffic on key thoroughfares. The data would be continuously updated. In this new world, people, more or less as hobbies, would tap into data of interest, develop a computer simulation of the real-world system, and visualize that simulation. The whole confluence would be a *mirror world.* One could imagine a mirror world on one's coffee table, embodied in a high-resolution 3-D display, showing the viewer the dynamic development of a municipality, or a town one is about to visit, or abstractly representing a distant political conflict, ... There could be many mirror worlds of interest to a given user; they could be traded around, personalized, improved. Mirror worlds could replace home aquariums (respectively econometric forecasts) as objects of affection (respectively concern). Like pets, we could have breeding of mirror worlds (improving characteristics by combining ideas from several mirror worlds). Like current IT software megapackages, we could have selling of mirror worlds for millions of dollars, if they addressed the needs of corporate decision makers.

It is assumed that the job of accessing the data, simulating the system, and visualizing the system would be very easy, because of some fundamental advances in how we do those tasks. So 'everyone' would want to be involved in the breeding and inspection of mirror worlds about issues of concern to them.

2.5 Compulsion to Data-Based Simulation

One aspect of this vision which should be faced is the compelling nature of visualization by computational simulation. Simply put, there is a big wow-factor in using the computer to simulate systems and present the results of simulation. This wow-factor can be dismissed by some mathematicians, committed to analysis and intellectual depth, as a superficial attribute, but vast numbers of others will be deeply moved by this vision.

Recent initiatives to build an e-cell [18] has attracted substantial amounts of attention from science journalists and science policy makers [7]. Can one build computer models which simulate the basic processes of life by building a mathematical model of the basic cycles in the cell and watching the evolution of what amounts to a system of coupled nonlinear ODE's? The question is open, but the demand cannot be doubted. The soon-to-be released MCell work [42] with its graphical evocation of molecular processes, seems likely to obtain an even broader response than e-cell, partly because of its evocation of spatial and stochastic components of cellular processes.

The appeal of such evocations will only increase with time. I expect to see that data-driven simulation efforts will gain tremendous attention and ultimately become ubiquitous in every field.

The current generation of teenagers has been raised by spending endless hours playing simulation games. A typical one, Soccer Manager 2000 is a data-rich simulation game, in which all the major teams and talents of the soccer world are available – the true teams, the true players, with faithful statistical characteristics. The game player analyzes the team and player and makes changes away from real team compositions and plays fantasy games.

As the generation weaned in this way matures and becomes consumers and decisions makers, we can expect that building and modifying data-driven simulations will become a ubiquitous aspect of normal life. This will create an amazing demand for data.

2.6 How Useful is all This?

One can easily make the case that we are gathering too much data already, and that fewer data would lead to better decisions and better lives [57]. But one also has to be very naive to imagine that such wistful concerns amount to much against the onslaught of the forces I have mentioned. Reiterating: throughout science, engineering, government administration, and business we are seeing major efforts to gather data into databases. Much of this is based, frankly, on blind faith, a kind of scientism, that feels that it is somehow intrinsically of worth to collect and manage data. In some cases commercial enterprises have made huge bets, assuming that knowledge of consumer web surfing clickstreams can be sold, traded or otherwise leveraged into value.

Similarly, giant investments have been made to decode the human genome and make it available to biological researchers. It is claimed that somehow this data will translate into an understanding of protein expression, and then to underlying biology and biochemistry.

We can't say at the moment whether such assumptions will prove valid or not. What we can say is that our society has chosen to gather massive amounts of data, that this trend is accelerating yearly, and that major efforts are underway to exploit large volumes of data.

3 Data Matrices

While data can be bewilderingly diverse, for the purposes of a talk like this one, we focus on a single uniform data structure, allowing us to describe a great number of applications with great economy.

We will consider what statisticians consider the usual data matrix, a rectangular array with N rows and D columns, the rows giving different *observations* or *individuals* and the columns giving different *attributes* or *variables.* In a classical setting we might have a study like the Framingham Heart study, with data gathered very carefully over decades, and ultimately consisting of about $N = 25,000$ records about the individual residents of Framingham Massachussets on $D = 100$ variables.

We now mention some recent examples we are aware of indicating the broad range of applications where we can have N by D data matrices.

3.1 Web Term-Document Data

Document retrieval by web searching has seen an explosive growth and acceptance over the last 5 years. One approach to document retrieval is the vector space model of information retrieval. In this model, one compiles *term-document matrices*, N by D arrays, where N, the number of documents, is in the millions, while D, the number of terms (words), is in the tens of thousands, and each entry in the array measures the frequency of occurrence of given terms in the given document, in a suitable normalization.

Each search request may be viewed as a vector of term frequencies, and the Matrix-Vector product of the Term-Document matrix by the search vector measures. [6, 45]

3.2 Sensor Array Data

In many fields we are seeing the use of sensor arrays generating vector-valued observations as a functions of time. For example, consider a problem in study of evoked potential analysis in neuroscience. An array of D sensors is attached to the scalp, with each sensor records N observations over a period of seconds, at a rate of X thousand samples, second. One hopes to use such data to witness the response of human neural system to various external stimuli. The array aspect allows one potentially to localize various effects within the head. [34].

3.3 Gene Expression Data

A very "hot" data analysis topic at the moment involves gene expression data. Here we obtain data on the relative abundance of D genes in each of N different cell lines. The details of how the experiment works are pointed to from [5]. The goal is to learn which genes are associated with the various diseases or other states associated with the cell lines.

3.4 Consumer Preferences Data

Recently on the world-wide-web we see the rise of attempts to gather information about browsing and shopping behavior of consumers – along with demographics and survey results – and to use this to modify presentation of information to users. Examples include recommendation systems like used at NetFlix and Amazon, and personalization systems like Xamplify. We mention briefly the NetFlix scheme http://www.netflix.com. Each consumer is asked to rate about 100 films; based on that rating, the consumer is compared to other customers with similar preferences, and predictions are made of other movies which might be of interest to the consumer based on experiences of other customers who viewed and rated those movies.

Here we have a rectangular array giving responses of N individuals on D movies, with N potentially in the millions and D in the hundreds (or eventually, thousands).

3.5 Consumer Financial History

Stine and Foster [56] give an example of credit card transaction records on 250,000 consumer/months, where several dozen variables (deomgraphics/payment history) are available on each consumer. Here N is 250,000 and D is in the thousands.

3.6 Tick-by-Tick Financial Data

In the past decade, there has been a radical improvement in the availability of high-frequency financial data, not just on stocks and bonds in say the US exhachanges, but on markets of all scales and locations. [15] is a book-length treatment describing treatment of high frequency tick-by-tick currency transaction data. One can imagine an N by D array with D variables giving a dozen or so exchange rates of foreign currencies to the dollar, while N time samples (with N large give the exchange rates on time scales of the very fine, extending perhaps for a very long period of time.

3.7 Imagery

We can view a database of images as an N-by-D data matrix. Each image gives rise to an observation; if the image is n by n, then we have $D = n^2$ variables. Different images are then our different individuals. We give an example in Figure

3.8 Hyperspectral Imagery

New optical sensors are able to obtain imagery not just with red-green-blue color sensitivity, (3 numbers per pixel) but instead a full spectrum with thousands of frequency bands being measured. Thus an image is not just, say, a 4096*4096 pixel array, but a 4096*4096*1024 pixel volume. Such data can be viewed as an N by D array. Suppose we have I images in our database, each of size n by n with S spectral bands. We can let $D = S$ and let $N = In^2$.

4 Data Analysis

In studying an N-by-D data matrix, we often refer to it as D-dimensional – because we take the view of N points in a D-dimensional space. In this section we describe a number of fundamental tasks of data analysis. Good references on some of these issues include [41, 51, 66]; I use these often in teaching.

4.1 Classification

In classification, one of the D variables is an indicator of class membership. Examples include: in a consumer financial data base, most of the variables measure consumer payment history, one of the variables indicates whether the consumer has declared Bankruptcy, the analyst would like to predict bankruptcy from credit history; in a hyperspectral image database all but one of the variables give spectral bands, an extra variable gives an indicator of ground truth chemical composition; the analyst would like to use the spectral band information to predict chemical composition.

Many approaches have been suggested for classification, ranging from identifying hyperplanes which partition the sample space into non-overlapping groups, to k-nearest neighbor classification; see [51].

4.2 Regression

In regression setting, one of the D variables is a quantitative response variable. The other variables are used to predict it. Examples include: in a financial data base, the variability of exchange rates today, given recent exchange rates; in a hyperspectral database an indicator of chemical composition. There is a well-known and widely used collection of tools for regression modeling; see [66, 22].

In linear regression modeling, we assume that the response depends on the predictors linearly,

$$X_{i,1} = a_0 + a_2 X_{i,2} + \ldots + a_D X_{i,D} + Z_i; \tag{1}$$

the idea goes back to Gauss, if not earlier. In nonlinear regression modeling, we assume that the response depends on the predictors in a general non linear fashion,

$$X_{i,1} = f(X_{i,2}, \ldots, X_{i,D}) + Z_i. \tag{2}$$

Linear regression modeling involves mostly linear algebra: the estimated coefficients of the least-squares method can be obtained by $\hat{a} = (X^T X)^{-1} X^T Y$, where Y is the column vector of response data. Nonlinear regression can involve: local linear fits, neural nets, radial basis functions, etc.

4.3 Latent Variables Analysis

In latent variables modeling we propose that

$$X = AS$$

where X is a vector-valued observable, S is a vector of unobserved latent variables, and A is a linear transformation converting one into the other. Often, the hope is that a few underlying latent variables are responsible for essentially the structure we see in the array X, and by uncovering those variables, we have achieved important insights.

Principal Component Analysis [28, 35, 38] is an early example of this. One takes the covariance matrix C of the observable X, obtains the eigenvectors, which will be orthogonal, bundles them as columns in an orthogonal matrix U and defines

$$S = U'X.$$

Hence we have the latent variable form with $A = U$.

This tool is widely used throughout data analysis in the sciences, engineering, and commercial applications. Projection on the space spanned first k eigenvectors of C gives the best rank k approximation to the vector X in a mean square sense.

A now standard application comes in latent semantic indexing, where it is used to perform web searching [6, 45]. One extends the PCA method to a singular value decomposition factorization

$$X = UDV'$$

where now V is the matrix of eigenvectors of C and D is the diagonal matrix with square roots of the eigenvalues of C. A query is a vector α indicating a list of terms to search for and responses are sorted based on values of

$$UD_kV'\alpha$$

to find documents with large query values, here D_k is a k-term approximation to the diagonal matrix D keeping only the k biggest terms. In effect the k-term approximation causes grouping of both terms and documents together, so that one can obtain 'hits' on documents that do not contain the precise term used, but that do contain a highly correlated term or terms.

PCA has been tried in image analysis, where it has been used to study images of faces. In that application, the eigenvectors can be viewed as images – "eigenfaces" – searching for matches of faces in a database of faces can then be processed in a fashion similar to the LSI model: if α gives the data vector for a new face, look for large entries in the output of the rank-k approximation for appropriate k, in

$$UD_kV'\alpha.$$

In the last decade, an important alternative to PCA has been developed: ICA – independent components analysis [13, 2, 11]. It is valuable when, for physical reasons, we really expect the model $X = AS$ to hold for an unknown A and a sparse or nonGaussian S. The matrix A need not be orthogonal.

An example where this occurs is with Array Data, where one assumes there are several sources, each one coupled with different strength to different sensors (for example, based on proximity of source to sensor). A typical example is the cocktail party problem: one has several microphones and several human speakers, the speakers are talking simultaneously and each microphone is picking up all the speakers at once.

In the talk we give an example based on EEG data by Jung et al. [34] from Terry Sejnowski's lab.

4.4 Clustering

Cluster Analysis could be considered a field all its own, part art form, part scientific undertaking. One seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar. There are many ways to do this, serving many distinct purposes, and so no unique best way.

An obvious application area would be in latent semantic indexing, where we might seek an arrangement of documents so that nearby documents are similar and an arrangement of terms so that nearby terms are similar. See for example [45].

Because we have mentioned gene expression data earlier, we briefly mention a figure presented earlier in the talk showing a gene expression array, (figure taken from [26]) while a figure based on a modern clustering method shows the array after suitable permutation of entries according to cluster analysis.

Recently, more quantitative approaches have been developed, of which we mention two here.

The first, Gene Shaving, is described in [26]; it has been developed by a team of statisticians and bioinformaticists,including my Stanford colleagues Hastie and Tibshirani. The underlying model is

$$X_{i,j} = \mu_0 + \sum_{k=1}^{K} \alpha_k \beta_k^T$$

where each β_k is a D-vector, each α_k is an N vector taking values 0 and 1, and in addition is sparse (relatively few 1's). An iterative, heuristic algorithm is used to fit layers $k = 1, 2, \ldots, K$ of the gene expression array.

The second, Plaid Modelling, [36] has been developed by my Stanford colleagues Lazzeroni and Owen. It seeks in addition to constrain each vector β_k to have entries either 0 and 1.

$$X_{i,j} = \mu_0 + \sum_{k=1}^{K} \mu_k \alpha_k \beta_k^T$$

Again an iterative, heuristic algorithm is used to fit layers of the gene expression array, layers $k = 1, 2, \ldots, K$, one at a time.

The two models differ in that plaid models have a complete 0-1 nature, giving a strict clustering form, while Gene Shaving clusters rows but does not constraint individual rows to have constant behavior.

5 High-Dimensionality

Our examples show that we are in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest. Our task is to find a needle in a haystack, teasing the relevant information out of a vast pile of glut.

This is a big break from the original assumptions behind many the tools being used in high-dimensional data analysis today. For many of those tools, it was assumed that one was dealing with a few well-chosen variables, for example, using scientific knowledge to measure just the right variables in advance.

But we shouldn't read anything pejorative into the fact that we don't know which variables to measure in advance. For example consider the functional data analysis case [49], where the data are curves (for example, in the hyperspectral imaging case where the data are spectra). The ideal variables might then turn out to be the position of certain peaks in those curves. What we measure automatically and reliably is the curves themselves, which we later can hope to analyze for peaks.

This *post-classical* world is different in many ways from the 'classical world'. The basic methodology which was used in the 'classical world' no longer is not strictly-speaking applicable. More or less, the theory underlying previous approaches to data analysis was based on the assumption of $D < N$, and $N \to \infty$. Many of the intellectually cleanest results concern properties of observations which were multivariate normal, and used extensively tools from linear algebra and from group theory to develop some exact distributional results. These results all fail if $D > N$. Even worse, they envision an asymptotic situation in which $N \to \infty$ with D fixed, and that also seems contradicted by reality, where we might even have D tending to ∞ with N remaining fixed.

The $D > N$ case is not anomalous; it is in some sense the generic case. For many types of event we can think of, we have the potential of a very large number of measurable quantifying that event, and a relatively few instances of that event. Examples include:

- Many genes, relatively few patients with a given genetic disease.
- Many samples of a persons' speech, relatively few speakers sampled.

It is in facing this intrinsic high dimensionality that I perceive there are great opportunities to make a contribution. I will now develop this theme more carefully.

In effect, we need to develop tools for the high dimensional case. These will often have a different spirit than in the past, as they will often be approximations, bounds, and asymptotics, whereas so much of classical multivariate analysis was beautifully exact (in the normal case). The new spirit will require different skills and will attract different kinds of scientists.

6 Curse of Dimensionality

6.1 Origins of the Phrase

The colorful phrase the 'curse of dimensionality' was apparently coined by Richard Belman in [3], in connection with the difficulty of optimization by exhaustive enumeration on product spaces. Bellman reminded us that, if we consider a cartesian grid of spacing 1/10 on the unit cube in 10 dimensions, we have 10^{10} points; if the cube in 20 dimensions was considered, we would have of course 10^{20} points. His interpretation: if our goal is to optimize a function over a continuous product domain of a few dozen variables by exhaustively searching a discrete search space defined by a crude discretization, we could easily be faced with the problem of making tens of trillions of evaluations of the function. Bellman argued that this curse precluded, under almost any computational scheme then foreseeable, the use of exhaustive enumeration strategies, and argued in favor of his method of dynamic programming.

We can identify classically several areas in which curse of dimensionality appears.

- In Optimization, Bellman's original usage. If we must approximately optimize a function of D variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^D$ evaluations on a grid in order to obtain an approximate minimizer within error ϵ .
- In Function Approximation. If we must approximate a function of D variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^D$ evaluations on a grid in order to obtain an approximation scheme with uniform approximation error ϵ .

- In Numerical Integration. If we must integrate a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^D$ evaluations on a grid in order to obtain an integration scheme with error ϵ .

The mathematics underlying these facts are all obvious; it is not the depth of the phenomenon that is noteworthy – for surely this is very superficial observation – but its ubiquity.

6.2 Curse in Statistical Estimation

Suppose we have a dataset with D variables, and we suppose that the first one is dependent on the others, through a model of the form.

$$X_{i,1} = f(X_{i,2}, ..., X_{i,D}) + noise_i.$$

Suppose that f is of unknown form, for example, we are not willing to specify a specific model for f, such as a linear model. Instead, we are willing to assume merely that f is a Lipschitz function of these variables and that $noise_i$ variables are in fact i.i.d. Gaussian with mean 0 and variance 1 .

How does the accuracy of estimation depend on N, the number of observations in our dataset? Let $\mathcal{F}$ be the functional class of all functions f which are Lipschitz on $[0,1]^d$. A now-standard calculation in minimax decision theory [30] shows that for any estimator $\hat{f}$ of any kind, we have

$$\sup_{f\in\mathcal{F}} E(\hat{f} - f(x))^2 \geq Const \cdot N^{-2/(2+D)}, \qquad n \to \infty.$$

This lower bound is nonasymptotic. How much data do we need in order to obtain an estimate of f accurate to within $\epsilon = .1$? using minimax decision theory gives us a way to answer this, and we obtain that trillions of samples are required.

The very slow rate of convergence in high dimensions is the ugly head of the curse of dimensionality.

7 Blessings of Dimensionality

Increases in dimensionality can often helpful to mathematical analysis. Typically, this is because of probability theory. The regularity of having many “identical” dimensions over which one can “average” is a fundamental tool.

7.1 Concentration of Measure

The “concentration of measure phenomenon” is a terminology introduced by V. Milman for a pervasive fact about probabilities on product spaces in high dimensions. Suppose we have a Lipschitz function f on the D-dimensional sphere. Place a uniform measure P on the sphere, and let X be a random variable distributed P Then

$$P\{|f(x) - Ef(x)| > t\} \leq C_1 \exp(-C_2 t^2). \tag{3}$$

where C_i are constants independent of f and *of dimension*. In short, *a Lipschitz function is nearly constant*. But even more importantly: the tails behave at worst like a scalar Gaussian random variable with absolutely controlled mean and variance.

This phenomenon is by no means restricted to the simple sphere case just mentioned. It is also true, in parallel form, for X taken from the multivariate Gaussian law with density

$$p(x) = (2\pi)^{-D/2} \exp(-\|x\|^2/2).$$

Variants of this phenomenon are known for many high-dimensional situations; e.g. discrete hypercubes $\mathbf{Z}_2^D$ and hamming distance. The roots are quite old: they go back to the isoperimetric problem of classical times. Milman credits the probabilist Paul Lévy with the first modern general recognition of the phenomenon. There is by now a vast literature on this

A typical example is the following. Suppose I take the maximum of D i.i.d. Gaussian random variables $X_1, \ldots, X_D$. As the maximum is a Lipschitz functional, we know from the concentration of measure principle that the distribution of the maximum behaves no worse than a standard normal distribution in the tails. By other arguments, we can see that the expected value of $\max(X_1, ..., X_D)$ is less than $\sqrt{2\log(D)}$. Hence the chance that this maximum exceeds $\sqrt{2\log(D)} + t$ decays very rapidly in t.

Another example is the following. Suppose I take the root-mean-square of D i.i.d. Gaussian random variables $X_1, \ldots, X_D$, or in simpler terms, the euclidean norm of the vector $X = (X_1, ..., X_D)$. As the norm is a Lipschitz functional, we know from the concentration of measure principle that again the distribution of the maximum behaves no worse than a standard normal distribution in the tails. By other arguments, we can see that the expected value of $||X||^2$ is D, so the expected value of $\|X\|$ is less than $\sqrt{D}$. Hence the chance that this norm exceeds $\sqrt{D} + t$ decays very rapidly in t.

7.2 Dimension Asymptotics

A second phenomenon, well-exploited in analysis, is the existence of results obtained by letting the number of dimension go to infinity. This is often a kind of refinement of the concentration of measure phenomenon, because often when there is a dimension-free bound like the concentration of measure phenomenon, there is a limit distribution for the underlying quantity, for example a normal distribution.

Return to the example of the maximum M_D of D i.i.d. Gaussian random variables. As remarked above, we know that the distribution of the maximum behaves no worse than a standard normal distribution in the tails. In fact, long ago Fisher and Tippett derived the limiting distribution, now called the extreme-value distribution []. That is, they showed that

$$Prob\{M_D - \sqrt{2\log(D)} > t\} \to G(t)$$

where $G(t) = e^{-e^{-t}}$.

Similarly, return to the example of the Euclidean norm N_D of D i.i.d. Gaussian random variables. Owing to the known properties of χ distributions,

$$Prob\{N_D - \sqrt{D} > t\} \to \Phi(t)$$

where $\Phi(t)$ is the standard Normal cumulative distribution function.

7.3 Approach to Continuum

Many times we have high-dimensional data because the underlying objects are really continuous-space or continuous-time phenomena: there is an underlying curve or image that we are sampling. Typical examples cited earlier include measurements of spectra, gaits, and images. Since the measured curves are continuous, there is a underlying compactness to the space of observed data which will be reflected by an approximate finite-dimensionality and an increasing simplicity of analysis for large D.

A classical example of this is as follows. Suppose we have d equispaced samples on an underlying curve $B(t)$ on the interval $[0, 1]$ which is a Brownian bridge. We have D-dimensional data $X_{i,D} = B(i/D)$, and discuss two computations where the large d behavior is easy to spot.

First, suppose we are interested in the maximum $\max_i X_{i,D}$. Then quite obviously, this tends, for large D to the random variable $\max_{t\in[0,1]} B(t)$, which has an exact distribution worked out by Kolmogorov and Smirnov.

Second, suppose we are interested in obtaining the principal components of the random vector. This involves taking the covariance matrix

$$C_{i,j} = Cov(X_i, X_j), 1 \leq i, j \leq D$$

and performing an eigenanalysis. On the other hand, the covariance kernel

$$\Gamma(s, t) = Cov(B(s), B(t)), \quad s, t \in [0, 1]$$

has the known form $\min(s, t) - ts$ and known eigenfunctions $\sin(\pi kt)$, for $k = 1, 2, \ldots$.

In this case, the first m eigenvalues of C tend in an appropriate sense to the first m eigenvalues of Γ and the eigenvectors of C are simply sampled sinusoids.

8 Exploiting the Blessings

We now give examples of how each of these blessings comes up in high-dimensional data analysis.

8.1 Model Selection

We begin with an example exploiting the concentration of measure phenomenon.

Suppose we have a linear regression problem, where there is a dependent variable $X_{i,1}$ which we want to model as a linear function of $X_{i,2}, ..., X_{i,D}$ as in (1).

However, D is very large, and let's suppose we're in a situation where there are thought to be only a few relevant variables, we just don't know which ones. If we leave many irrelevant variables in the model, we can easily get very poor performance. For this reason, statisticians have, for a long time, considered model selection by searching among subsets of the possible explanatory variables, trying to find just a few variables among the many which will adequately explain the dependent variable. The history of this approach goes back to the early 1960's when computers began to be used for data analysis and automatic variable selection became a distinct possibility.

We have to be on guard to the problem of what was classically called data mining – over-optimistic assessments derived by searching through noise and fooling ourselves that we have found structure.

One approach, used since the 1970's, is to optimize over subset models the complexity penalized form

$$\min \text{RSS}(\text{Model}) + \lambda \text{Model Complexity},$$

where RSS denotes the residual sum of squares of the residuals $X_{i,1} - Model_{i,1}$, and the model complexity is the number of variables $X_{i,2}, \dots, X_{i,D}$ used in forming the model. Early formulations used $\lambda = 2 \cdot \sigma^2$, where σ^2 is the assumed variance of the noise in (1). The overall idea is to impose a cost on large complex models.

More recently, one sees proposals of the form $\lambda = 2 \cdot \sigma^2 \cdot \log(D)$. With these logarithmic penalties, one takes into account in an appropriate way the true effects of searching for variables to be included among many variables. A variety of results indicated that this form of logarithmic penalty is both necessary and sufficient, for a survey see [31]. That is, with this logarithmic penalty, one can mine one's data to one's taste, while controlling the risk of finding spurious structure.

The form of the logarithmic penalty is quite fortunate. The logarithm increase quite slowly with D – a faster increase would indicate that automatic variable selection is in general hopeless: one loses too much by searching for the right variables. And, interestingly, the logarithm is directly due to the concentration of measure phenomenon. That is to say, the presence of the exponential decay in the concentration of measure estimates (3) is ultimately responsible for the logarithmic form of the penalty.

8.2 Asymptotics for Principal Components

We now turn to an example of our second "blessing of dimensionality" – that results for high dimensions can be easier to derive than for moderate dimensions.

Now suppose we have data $X_{i,j}$ where the vectors $X^{(i)} = (X_{i,j} : 1 \leq j \leq D)$ are assumed samples from a Gaussian distribution with mean zero and covariance Γ. We are interested in knowing whether $\Gamma = I$ as compared to $\Gamma \neq I$. Depending on our alternative hypothesis, it might be very natural to rephrase our question as $\lambda_1 = 1$ versus $\lambda_1 > 1$, where λ_1 is the top eigenvalue of the covariance matrix. It then becomes natural to develop a test based on l_1, the top eigenvalue of the empirical covariance matrix $C = N^{-1}X'X$. it then becomes important to know the null distribution of l_1. Exact formulas go back to work by Ted Anderson of Stanford of the 1950's published in Anderson (1963), but are not very useful in the setting we consider; they cover the D fixed $N \to \infty$ case. Already for moderate D, and N proportional to D one cannot really apply them.

Suppose instead we are in a setting of many observations and many variables. What is the behavior of the top eigenvalue of $C_{D,N}$? Consider a sequence of problems where $D/N \to \beta$ – large dimension, large sample size.

This is a problem which has been studied for decades; see [32] for references. Classical results in random matrix theory – in the spirit of the Wigner semicircle law – study infinite matrices, and in accord with our "second blessing of dimensionality", give information about the bulk spectrum of C; but unfortunately they do not accurately predict the very top eigenvalue.

Recently, Tracy and Widom have made very substantial innovations in the study of the topmost eigenvalues of certain ensembles of infinite random matrices – so called Gaussian Unitary and Orthogonal Ensembles. Iain Johnstone recognized that this work had applications in a statistical setting, and building also on the work of Kurt Johansson has been able to obtain asymptotic results for the top eigenvalue of $C_{D,N}$ in the so-called null case – results which are relatively easy to apply in practice.

Moreover, empirical studies show that the results, though derived asymptotically, are in fact useful for d as small as 5. Hence we obtain, from a high-dimensional analysis, useful results in moderate dimensions.

8.3 Fourth-Order Structure of Stochastic Processes

We now consider an example of our third blessing of dimensionality – how in a setting where the data are curves, continuum theory may furnish an interpretation of the results.

Consider a collection of underlying functions $f(t, \omega)$ defined on the index set $T \equiv [0, 1]$, with parameter ω chosen from $[0, 1]$, defined by

$$f(t;\omega) = \begin{cases} t & t < \omega \\ t-1 & t \geq \omega \end{cases}$$

Each curve very simple behavior: it jumps down by 1 at the jump time ω; otherwise it just increases at unit slope. It is a model of a singularity occurring at any of a range of times.

Consider a simple numerical experiment. With $N = D = 32$ we define vectors Y_i, $i = 1, ..., D$, each one a simple digitization of *Ramp*.

$$Y_i(t) = f(t/n, i/n), \quad 1 \leq t \leq N; \quad 1 \leq i \leq D.$$

The database $\mathcal{Y}$ thus consists of 32 signal patches, and we use this database as input to the so-called JADE procedure. That is, we calculate from this data the empirical $(32)^4$ cumulant tensor, and we use JADE to attempt a diagonalization of this tensor. The result will be an orthonormal basis with 32 elements depicted in a Figure to be shown in the talk. The structure of the basis is rather remarkable; it has many of the features of a wavelet basis.

- *Dyadic Scales.* The elements seem visually to posses a variety of scales; they can be arranged in a dyadic pyramid, with 16 elements at the finest scale, 8 elements at the next finest scale, and so on.

- *Translations.* Within one scale, the elements seem to be approximately translates of each other, so that (at fine scales particularly) there are elements located roughly at positions $t_{j,k} = k/2^j$.

- *Cancelation.* The elements at fine scales seem to be oscillatory, with two vanishing moments.

In my Talk I will present a Figure showing a Daubechies nearly-symmetric basis with 6 vanishing moments. Another Figure will gives a few side-by-side comparisons between these “JADElets” and certain Daubechies nearly-symmetric wavelets. The reader may notice a striking resemblance.

In short, instead of the almost-diagonalizing basis being arbitrary, it resembles wavelets, an object from the continuum theory. Presumably, this resemblance becomes stronger with large D, and so the interpretation becomes increasingly simple. If so, we are blessed by the approach to an underlying continuum model.

9 Predictions

At this point, I hope to have convinced the reader of three major intellectual and societal trends:

- The Ubiquity of Data

- The Importance of Data Analysis

- The Pervasiveness of High-Dimensionality

I believe that the future of data analysis is now in the hands of those who would explore the high-dimensional situation, facing the situation as it stands now, at the turn of the century.

I hope also to have made clear the basic aspects of the high-dimensional situation

- Three aspects of the curse of dimensionality

- Three blessings of dimensionality

The occasion seems to demand that at this point I "deliver the goods": that I state some number of open problems (23?), leaving open to future generations to solve. I think this would be really foolhardy on my part.

Instead, I will mention some different directions in which I expect to see much further progress...

9.1 High-Dimensional Approximate Linear Algebra

9.1.1 Approximations through Randomness

Modern computational linear algebra is a great success story. It is currently used to solve massive data processing problems - fitting gigantic models, scheduling fleets of airliners,

Nevertheless, in a certain sense it is slow. Inverting a matrix takes $O(N^3)$ operations, which for large N (say in the millions) is prohibitively expensive. The Strassen algorithm reduces this to N^s with s about 2.7, but that is still heavy. Hence, we are entitled to dream that these fundamental components can be sped-up.

I will mention two recent articles that give an idea that something can be done

Owen [48] considered the problem of determining if a dataset satisfied an approximate linear relation of the form (1). With $N > D$ observations, the usual method of assessing linearity would require order ND^3 operations, which could be a computationally heavy burden for D large. Instead, Owen shows how to use randomized algorithms to assess linearity in order $N^{2/3}$ operations.

Frieze, Kannan, and Vempala [20] considered the problem of obtaining a rank k approximation to a data matrix X of size N by D. With $N > D$ observations, the usual method of singular value decomposition would require order ND^3 operations, which could again be a computationally heavy burden for D large. Instead, Frieze, Kannan, and Vempala show how to use randomized algorithms to obtain an approximate description of the rank k approximation in a number of operations independent of N and D, and polynomial in k and in the error tolerances.

In these cases, it is the phenomenon of concentration of measure that we are exploiting. However, in the cited papers, we are using only the most elementary form of the phenomenon. With the growing importance of massive databases, results like these seem to be increasingly pertinent and attractive. Supposing that these develop as I expect, increasingly sophisticated versions of concentration of measure would develop and be applied.

9.1.2 Approximation through Computational Harmonic Analysis

There is a different sense in the term approximate linear algebra that seems destined to become important. This derives from our third blessing: approach to the continuum. In recent years, the discipline of computational harmonic analysis has developed large collections of bases and frames such as wavelets, wavelet packets, cosine packets, Wilson bases, brushlets, and so on [16, 39]. As we have seen, in two cases, that a basis being sought numerically by a procedure like principal components analysis or independent components analysis, will resemble some previously-known basis deriving from continuum theory. We saw, in fact, that the covariance of a stationary process is almost diagonalized by the Fourier basis, and that the fourth cumulant tensor of a process with discontinuities is almost diagonalized by the Wavelet basis.

There are numerous other examples where a fixed basis known in harmonic analysis does as well as a general procedure. The article [17] shows that in recognizing facial gestures, the Gabor basis gives better classification than techniques based on general multivariate analysis – such as Fisher scores and Principal component scores.

The expectation one can develop from such evidence is that when the data are curves or images, instead of using methods which search for an arbitrary basis or subspace, we should look for it in a pre-constructed set. There is a good statistical reason for this. In effect, a basis can be a hard thing to estimate well – there is substantial statistical uncertainty even when N is large. If we discover that a certain known basis is almost diagonalizing, we might use that basis, acting as if it were exactly diagonalizing, and avoiding thereby a substantial component of estimation error.

In an extreme case, this is easy to see. Suppose we have a dataset with $N = 1$ and D very large. Ordinarily, in this high-dimensional case we can do nothing: we have 1 observation! But if the data are obtained from one realization of a stationary Gaussian stochastic process, we can in fact do something: we have all the apparatus of modern time series analysis at our disposal, and we can indeed estimate the spectrum, and learn the full probability distribution. In effect, spectrum estimation is relying heavily on the fact that we know the covariance to be almost diagonal in the Fourier basis.

The article [40] develops a full machinery based on this insight. A dataset is analyzed to see which out of a massive library of bases comes closest to diagonalizing its empirical covariance, the best basis in the library is constructed, and the data are processed using that basis.

9.2 High Dimensional Approximation Theory

We now consider the curse of dimensionality itself, and re-examine the basic phenomenon mentioned earlier.

The key assumption that makes it hard to approximate a function of D-variables is that f may be an arbitrary Lipschitz function. With different assumptions, we could have entirely different results. *Perhaps there is a whole different set of notions of high-dimensional approximation theory, where we make different regularity assumptions and get very different picture.*

This point is underlined by a result of Andrew Barron [1]. Let $\mathcal{F}L^1$ denote the collection of functions with Fourier transforms in L^1. Consider the class $\mathcal{F}$ of functions of $\mathbf{R}^D$ with

$$\nabla f \in \mathcal{F}L^1. \tag{4}$$

Normally, in approximation results, one expects that objects of smoothness s can be approximated at rate $n^{-s/D}$ by n-term approximations (e.g. polynomial, trigonometric, ...).

Apparently, (4) is a condition on the first derivative of f. Hence one expects that the condition (4) leads to an approximation rate $O(n^{-1/D})$, which is very bad in high dimensions. In fact Barron showed that functions of this type can be approximated at rate $O(n^{-1/2})$ independent of dimension.

This result made quite an impression, and was paraphrased by many as saying that Barron had “cracked the curse of dimensionality”. In fact, now the phenomenon is better understood, and we know that there are many functional classes which allow one to evade the curse of dimensionality.

The phenomenon is easy to understand using ideas and terminology from harmonic analysis. Consider the class of functions $\mathcal{F}(M)$ representable integral representation $f = \int A(x;t)\mu(dt)$ with $\int |\mu(dt)| \leq M$ where each $A(\cdot;t)$ is a bounded functions bounded by 1. We call this an L^1 combination of L^∞ atoms. Then owing to a simple soft argument dating back to B. Maurey in geometry of Banach spaces and S.B. Stechkin in Fourier analysis, there is an m-term sum $f_m = \sum_j a_j A(\dot{;} t_j)$ with sup-norm error $|f - f_m|_\infty \leq C \cdot m^{-1/2}$.

Niyogi and Girosi [46] have pointed out that if we consider the set $\mathcal{F}(m, \text{Gaussians})$ generated with parameter $t = (x_0, s)$ and $A(x;t) = \exp(-\|x - x_0\|^2/s^2)$, so that f is a superposition of gaussian bumps, then we equivalently avoid the curse of dimensionality by using the Radial Basis function approximation scheme. The condition is that the sum of the heights of all the bumps is at most a constant M, with no restriction whatever on the width or position of the bumps.

Another charming example is $\mathcal{F}(M, \text{Orthants})$ uses the parameter set of shifted orthants. We let $t = (x_0, k)$ where k is an orthant indicator, and let $A(x;t)$ be the indicator of orthant k with apex at x_0. Then again, if the integral is at most M we obtain an approximation rate $O(m^{-1/2})$ independent of dimension. An example of a function satisfying the condition is a cumulative distribution function in $\mathbf{R}^D$ – in which case the result is well known under the guise of the monte-carlo method. More generally, consider any superposition of 2^D functions, each orthantwise monotone for a different orthant.

We can easily see derive equivalent conditions. For example, the class of superpositions of Gaussian bumps studied by Niyogi and Girosi is simple Yves Meyer's bump algebra [43]; this corresponds to a ball in a Besov space $B^D_{1,1}(\mathbf{R}^D)$, so that the functions in $\mathcal{F}(M, \text{Gaussians})$ are getting increasingly smooth in high dimensions [23, 43]. In short, one does not really crack the curse of dimensionality in Barron's sense; one simply works with changing amounts of smoothness in different dimensions. The ratio S/D between the smoothness degree and the dimension stays constant, and hence the $N^{-S/D}$ result does not become punishing.

Similarly, for the shifted orthants example, the condition of integral representation with $\int |\mu(dt)| \leq 1$ is the same as the condition that the mixed derivative

$$\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_D} f \in L^1. \tag{5}$$

In analogy to $D = 1$ where this reduces essentially to bounded variation, we might call this a condition of bounded *mixed variation*. The functions in the relevant set $\mathcal{F}(M, \text{Orthants})$ are getting increasingly smooth in high dimensions. One does not really conquer the curse of dimensionality; one simply works with changing amounts of smoothness in different dimensions.

Or perhaps we should reconsider this assessment. The functions in $\mathcal{F}(M, \text{Gaussians})$ are minimally continuous; the functions in $\mathcal{F}(M, \text{Orthants})$ are not even continuous. In each case, the usual ways of measuring smoothness suggest that functions in these classes have D weak derivatives. It would be more intuitive had this calculation turned out differently, and we agreed that these functions don't have many derivatives. Perhaps we need a different way to measure degree of differentiability, in which discontinuous functions are not claimed to possess large numbers of derivatives.

R.R. Coifman and J.O. Stromberg [12] have gone in exactly this direction. Viewing the condition of bounded mixed variation as a natural *low order* smoothness condition, they have explored its consequences. Roughly speaking, functions obeying the BMV condition have the character that they are built up from indicators of dyadic rectangles, that very few rectangles are needed, that the ones which are needed tend be very long in all directions except for one. In effect, locally the function varies only in a single direction at most x, and so it becomes understandable that it can be estimated at a rate essentially independent of dimension.

The argument for studying approximation theory for spaces of mixed variation then becomes:

- The smoothness condition is in fact weak.
- The class of functions obeying the condition are in fact quite interesting.
- The algorithm obtaining the estimation rate is new and interesting.

These brief remarks may suggest to the reader that substantial opportunities for interesting results over the coming years in high-dimensional approximations.

10 Some Personal Experiences

Why does it make sense to take time on the program at a major mathematical meeting to discuss data analysis? Ultimately, the answer is a personal one: from my own research experiences, I have found that by understanding what is happening in the field of data analysis, one poses interesting questions in mathematics per se. Extrapolating from my own experience suggests this may hold more broadly.

10.1 Interpretation of High-Dimensional Data Analysis

As high-dimensional data analysis progresses to settings where the data are images or curves, we have predicted above that there will be a substantial role for harmonic analysis to understand and interpret the results. But sometimes, the results will not be known in harmonic analysis, and the interpretation may even call for new constructions in harmonic analysis.

As an example, consider the following sort of image analysis problem which has attracted a great deal of attention recently.

We gather naturally-occurring images, extract image patches, assign unravel each patch to become a row of our data array, and create a data array with the image patches embedded in it. We then perform independent components analysis or sparse components analysis.

The result will be a basis of imagelets (2-d image patches) or movielets (if we were analyzing 3-d image patches). Interesting examples of this work include work by Olshausen and Field [47], Bell and Sejnowski [2] van Hateren and Ruderman [25], and Donato et al. [17].

Recall that principal components of image data typically produce sinusoidal patterns. ICA in contrast, typically produces basis functions unlike classical bases.

In the talk I will present an example of a movie created by van Hateren and Ruderman.

In these cases, because the data are images, we may expect that the result of high-dimensional data analysis – a basis – should be a basis which corresponds to a basis for the continuum. (Recall our earlier examples of this).

The sought-for continuum basis would consist of anisotropic elements at a range of scales, locations, orientations, and length/width relations. These properties are not available from classical constructions, such as wavelets.

In short, it seems as if one can discover phenomena by high-dimensional data analysis which ought to have been built by harmonic analysts, but which were not.

At the moment, it appears that these empirical results can best be explained in terms of two recent systems.

- Ridgelets [8]. One can build a frame or a basis consisting of highly directional elements, roughly speaking ridge functions with a wavelet profile. With a and b scalars and u a unit vector (direction) the simplest ridgelets take the form

$$\rho(x; a, b, u) = \psi((u'x - b)/a)/a^{1/2}.$$

- Curvelets [9]. One can build a frame consisting of highly directional elements, where the width and length are in relation $width = length^2$. The construction is a kind of elaborate multiresolution deployment of ridgelets.

These new systems of harmonic analysis offer a series of interesting mathematical opportunities – for example in the analysis of singularities along curves [10]. They also stand in interesting contrast to existing tools of microlocal analysis such as the tool Eli Stein calls "Second Dyadic Decomposition".

In general there is good reason to expect high-dimensional data analysis in th efuture to offer challenges to harmonic analysis. Both are concerned with finding almost-diagonal representations of operators. However the operators that arise in data analysis are continuously changing, and new phenomena can arise each time a new kind of sensor or data source comes on line. Hence, interpretation of the results of data analysis may continually suggest new schemes of harmonic analysis.

10.2 Geometric Structures in Low-Dimensional Data Analysis

One reason that high-dimensional data analysis should be mentioned at this conference is that in fact we know so little. At recent rates of progress, there is more than enough work to last through the coming century.

My evidence for this belief comes from our *meager understanding of data representation even in two and three dimensions.*

In my Talk I will give a few figures illustrating the problem of detecting Filaments in noisy data. This problem is of current interest because of the imminent arrival of data from the XMM satellite observatory, which will shed light on inhomogeneities in the distribution of matter at the very earliest detectable moments in the universe [19, 52].

I hope also to present figures from work by Arne Stoschek, illustrating the problem of recovering filamentary structure in 3-D biological data.

A third set of figures may suggest the problem of finding a curve embedded in extremely noisy data.

These problems suggest the relevance of work in harmonic analysis by Peter Jones [33] and later by Guy David and Stephen Semmes [14].

Jones' traveling salesman theorem considers the problem: given a discrete set S of points in $[0,1]^2$, when is there a rectifiable curve passing through the points? His theorem says that this can be decided solely on the basis of certain local functionals. If we let Q denote a dyadic subsquare of $[0,1]^2$, and define t_Q to be the thickness of the thinnest parallel strip containing all the points in the dilate $S \cap 3Q$, then defining $\beta_Q = t_Q/\ell(Q)$, where ℓ is the sidelength of Q. This is a measure of how anisotropic the set S is at the scale $\ell(Q)$ – a sort of width/length ratio. If the set lies on a curve, this could be very small. Jones proved that the finiteness of the sum $\sum \beta(Q)^2 \ell(Q)^2$ controls the existence of a rectifiable curve.

David and Semmes have studied the problem of approximations in R^D by k-dimensional varieties, and have developed machinery for this case which specializes to yield the original Jones result.

In view of the importance for data analysis of k-dimensional linear approximation in R^D – this is, after all, the problem of principal components analysis – it seems possible that these tools will one day be routinely applied not just in analysis but also in high-dimensional data analysis. The Yale thesis of Gilad Lerman, under the direction of Jones and R.R. Coifman, has made an auspicious start in that direction.

In recent work with Xiaoming Huo, I have been developing some computational tools which are related to this work. We define the Beamlets a dyadic multiresolution family of line segments, the beamlet graph, and beamlet transform – the family of line integrals along beamlets. The idea is to gather data over a special family of line segments. We develop beamlet-based algorithms for finding filaments in noisy data and recognizing objects with curved boundaries in noisy data. In the Talk, I will show figures of some preliminary results.

In work with Ofer Levi, we have made preliminary efforts with beamlets in $\mathbf{R}^3$, which is of obvious interest for structural biology and medical imaging. In the Talk, I will show figures of some of these.

It seems to me that we are just at the beginning stages of having tools for characterizing data around geometric structures in 2- and 3- dimensions. Modern data analysis has an urgent need for more tools, and I suspect that modern harmonic analysis can make progress and be a major contributor in that effort.

11 Conclusion

One of Tukey's implicit points was that data analysis would often be a relatively elementary activity, for example conducted by hand, as in the stem-and-leaf plots and boxplots developed in his book *Exploratory Data Analysis*.

Surely this is an eccentric position for Tukey to have invested with so much time and energy. Tukey was one of the nation's leading intellectuals; the President of Princeton in the 1970's, William F. Bowen, once told me that he regarded Tukey as a "National Treasure". I suspect that Tukey felt it was important to develop stem-and-leaf and similar ideas partly because the utter simplicity of the ideas would underscore the separation between data analysis and mathematics.

This aspect of Tukey's position has dated the most rapidly. Data analysis today is not an unsophisticated activity carried out by hand; it is much more ambitious, and, in my opinion, an intellectual force to be reckoned with. The examples of Olshausen and Field and van Hateren and Ruderman shows that in fact data analysis is now producing quite sophisticated objects – bases for high-dimensional spaces with rich and mathematically unprecedented structure. These objects are complex enough to provide challenges to mathematicians – "Why doesn't mathematics currently provide a good language for describing what I am seeing in my experiments?" and "Can't you formally develop systems of harmonic analysis which look like this?" I was quite elated when I encountered this work; it revealed to me that data analysis has now reached a point of sophistication where it can challenge and enrich mathematical discourse.

In 1900 Hilbert closed his paper with the question of whether Mathematics would have a schism:

> ... the question is urged upon us whether mathematics is doomed to the fate of those other sciences that have split up into separate branches, whose representatives scarcely understand one another and whose connection becomes ever more loose.

In fact we have seen in the past century exactly the sort of schism Hilbert worried about, and our talk today gives an example of a schismatic movement.

In Hilbert's closing, he asserted that Mathematics could avoid schism, saying

> ...the farther a mathematical theory is developed ... unexpected relations are disclosed between hitherto separate branches of the science ... its organic character is not lost but only manifests itself with clarity.

My personal research experiences, cited above, convince me of Hilbert's position, as a long run proposition, operating on the scale of centuries rather than decades.

References

[1] Barron, A. (1993) Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Tr. Information Theory*, **39**, 3, 930-945.

[2] Bell, A.J. and Sejnowski, T.J. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* **7** 1129-1159.

[3] Bellman, Richard (1961) *Adaptive Control Processes: A Guided Tour.* Princeton University Press.

[4] Bellman, Richard (1984) *Eye of the Hurricane: an autobiography.* Singapore: World Scientific.

[5] Bassett, D.E., Eisen, M.B., Boguski, M.S. Gene expression informatics – it's all in your mine. *Nature Genetics Supplement* **21**, 51-55.

[6] Michael W. Berry and Susan T. Dumais, and Gavin W. O'Brien (1995) Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review* **37**:4, pp. 573-595.

[7] Declan Butler (1999) Computing 2010: from black holes to biology *Nature*, **402** pp C67-70. Dec 2, 1999.

[8] Candès, E. and Donoho, D. (1999). Ridgelets: the key to high-dimensional intermittency?. *Phil. Trans. R. Soc. Lond. A.* **357** 2495-2509.

[9] Candès, E. J. and Donoho, D. L. (2000). Curvelets: a surpisingly effective nonadaptive representation of objects with edges. in *Curve and Surface Fitting: Saint-Malo 1999* Albert Cohen, Christophe Rabut, and Larry L. Schumaker (eds.) Vanderbilt University Press, Nashville, TN. ISBN 0-8265-1357-3

[10] Candès, E. J. and Donoho, D. L. (2000). Curvelets and Curvilinear Integrals. Technical Report, Department of Statistics, Stanford University.

[11] Cardoso, J.F. and Souloumiac, A. (1993) Blind Beamforming for non-Gaussian Signals. *IEE Proceedings-F.* **140** 352-370.

[12] R.R. Coifman, "Challenges in Analysis I", address at *Mathematical Visions towards the year 2000*, Tel Aviv University, August 1999.

[13] Comon, P. (1994) Independent Component Analysis, a new concept? *Signal Processing* **36** 287-314.

[14] G. David and S. Semmes (1993) *Analysis of and on Uniformly Rectifiable Sets.* Math Surveys and Monographs **38**, Providence: AMS.

[15] Dunis, Christian and Zhou, Bin (eds.) (1998) Nonlinear Modelling of High-Frequency Financial Time Series. J. Wiley: New York.

[16] Donoho, D.L., Vetterli, M., DeVore, R.A., and Daubechies, I. (1998) Data Compression and Harmonic Analysis. *IEEE Trans. Info. Thry.* **44**, 6, 2435-2476.

[17] Gianluca Donato, Matian Bartlett, Joseph Hager, Paul Ekman, and Terrence J. Sejnowski. (1999) Classifying Facial Actions. *IEEE Trans. Pattern Anal and Machine Intell.* **21**, 974-989.

[18] E-Cell: http://www.e-cell.org

[19] Frenk, C.S., White, S.D.M., Davis, M., Efstathiou, G. (1988) The formation of dark halos in auniverse dominated by cold dark matter. *Astrophys. J.* **327** 507-525.

[20] Frieze, Alan, Kannan, Ravi, Vempala, Santosh (1998) Fast Monte-Carlo Algorithms for Finidng Low-rank Approximations. Technical Report, Yale University Department of Computer Science.

[21] Friedman, Jerome H., Stuetzle, Werner (1981) Projection pursuit regression. *Journal of the American Statistical Association*, **76** 817-823.

[22] Jerome Friedman, Trevor Hastie and Robert Tibshirani (2001) *ELEMENTS OF STATISTICAL LEARNING: Prediction, Inference and Data Mining* Springer: New York

[23] Frazier, M., and B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*, NSF-CBMS Regional Conf. Ser. in Mathematics, Vol 79, American Math. Soc., Providence, RI, 1991.

[24] Ivor Grattan-Guiness A sideways look at Hilbert's Twenty-Three Problems of 1900. Notices AMS **47** 752-757.

[25] van Hateren, J.H. and Ruderman, D.L. (1998) Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond.* B **265**

[26] Trevor Hastie, Robert Tibshirani, Michael Eisen, Pat Brown, Doug Ross, Uwe Scherf, John Weinstein, Ash Alizadeh, Louis Staudt, David Botstein, Gene Shaving: a New Class of Clustering Methods for Expression Arrays, Technical Report, Department of Statistics, Stanford University, January, 2000.

[27] D. Hilbert, (1902) Mathematical Problems, *Bull. Amer. Math. Soc.* 8 (1902), 437-79; English Translation of article in *Arch. Math. Physik* **1** 44-63,213-237.

[28] Hotelling, H. (1933) Analysis of complex statistical variables into principal components. Journel of Educational Psychology 24, 417-441, 498-520.

[29] Hotelling, H. (1939) Tubes and spheres in n-spaces, and a class of statistical problems. *Amer. J. Math.* **61**, 440-460.

[30] Ibragimov, I. and Khasminskii, R.Z. (1981) Statistical estimation: asymptotic theory. New York: Springer.

[31] Johnstone, Iain (1998) Oracle Inequalities and Nonparametric Functional Estimation in *Documenta Mathematica ICM 1998* **III**, 267-278.

[32] Johnstone, Iain (2000) On the distribution of the largest principal component. Technical Report, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~imj/Reports/2000/largepc.ps

[33] P. W. Jones (1990) "Rectifiable Sets and the Travelling Salesman Problem." *Inventiones Mathematicae*, **102** 1-15.

[34] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin Mckeown, Vicente Iragui, Terrence Sejnowksi (2000) Removing Electroencephalographic Artifacts by blind source separation. *Psychophysiology* **37** 163-178.

[35] Kari Karhunen (1947) Uber Lineare Methoden in Wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fenn. 37.

[36] Lazzeroni, Laura, and Owen, Arthur. Plaid Models for Gene Expression Data Technical Report, Department of Statistics, Stanford University, March, 2000.

[37] David Leonhardt, "John Tukey, 85, Statistician; Coined the Word 'Software', " *New York Times*, July 28, 2000.

[38] Michel Loève. Fonctions aleatoires de second ordre. C. R. Acad. Sci. 220 (1945), 222 (1946); Rev. Sci. 83 (1945), 84 (1946).

[39] S.G. Mallat. (1998) *A Wavelet Tour of Signal Processing*. Academic Press.

[40] S.G. Mallat, G. Papanicolaou, and Z. Zhang. Adaptive covariance estimation of locally stationary processes. *Ann. Statist.* , 26, February 1998.

[41] K. V. Mardia, J. T. Kent, J. M. Bibby. *Multivariate analysis* , London ; New York : Academic Press, 1979.

[42] MCell: see http://www.mcell.cnl.salk.edu

[43] Meyer, Y., *Wavelets and Operators*, Cambridge University Press, 1992.

[44] Vitaly Milman. The Heritage of P. Levy in Geometrical Functional-Analysis Asterisque, **157** pp. 273-301

[45] Murtaugh, F., Starck, J.-L., Berry, M.W. (2000) Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *Computer Journal* **43**, pp. 107-120.

[46] P. Niyogi and F. Girosi (1998) Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics* **10** pp. 51-80.

[47] Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607-609.

[48] Owen, Arthur. Assessing Linearity in High Dimensions, Technical Report, Department of Statistics, Stanford University, 1999.

[49] Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis.* Springer: New York.

[50] Constance Reid (1970) *Hilbert*, Springer-Verlag, New York.

[51] Ripley, Brian D. (1996). *Pattern recognition and neural networks.* New York: Cambridge University Press, 1996.

[52] Sahni, V., Sathyaprakash, B.S., Shandarin, S.F. (1994) The evolution of voids in the adhesionapproximation. *Astrophys. J.* **431** 20-40.

[53] *Scientific American*, July 2000. Special Section: The BioInformatics Industry.

[54] C.A. Scott (1900) The International Congress of Mathematicians in Paris, *Bull. Amer. Math. Soc.* 7 (1900), 57-79

[55] David W. Scott (1996) Multivariate Density Estimation

[56] Bob Stine and Dean Foster. (1999) Variable Selection in Credit Modelling. http://www-stat.wharton.upenn.edu/~bob/research/baltimore.pdf

[57] David Shenk (1998) *Data Smog: Surviving the Information Glut.* Harper, San Francisco.

[58] V. Pestov (2000) On the geometry of similarity search: dimensionality curse and concentration of measure. *Information Processing Letters* **73** 47-51.

[59] C.A. Tracy and H. Widom (1996) On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics* **177** 727-754.

[60] Lionel Trilling (1953) Preface in *Homage to Catalonia* by George Orwell. 1987 Paperback: Harcourt Brace.

[61] Edward R. Tufte *The Visual Display of Quantitative Information* Cheshire, Conn: Graphics Press, 1983.

[62] Tukey, John W. (1962) The future of data analysis, *Ann. Math. Statist.*, **33**, 1–67. MR24:A3761, I. J. Good.

[63] John W. Tukey and Frederick Mosteller (1968) Data Analysis, Including Statistics in *Handbook of Social Psychology*, Gardner Lindzey and Elliot Aronson, eds. 80-112.

[64] John W. Tukey (1977) Exploratory Data Analysis. Addison Wesley.

[65] Tukey, JW. (1985) Personal Correspondence with DLD.

[66] W.N. Venables, B.D. Ripley. *Modern applied statistics with S-PLUS*, 3rd ed. New York: Springer, 1999.

[67] Howard Wainer *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonarparte to Ross Perot* Copernicus books, Springer, NY: 1997

[68] Weyl, H. (1939) On the volume of tubes. *Amer. J. Math.* **61**, 461-472